07: Transformations for Regression and Residual Plots

Stat 120 | Fall 2025

Prof Amanda Luby

Kathleen Vongasthorn (2007) compiled data on the top 100 films released in 2004 with the highest domestic gross over the course of its theatrical release (in US dollars). The variables in her dataset (called `Movies2004.csv`) are:

- `screens`: total number of screens on which these films played in the US
- `opening`: total profit made during opening weekend in the US
- `domestic`: total domestic gross
- `worldwide`: total worldwide gross
- `oscar`: whether the film was nominated for an Academy Award
- `globe`: whether the film was nominated for a Golden Globe

0. Load libraries and data. Check that you can view the data and that all the variables mentioned above are there. What does each row represent?

```
library(tidyverse)
library(broom)
library(patchwork)
Movies2004 = read_csv("http://math.carleton.edu/Stat120/RLabManual/Movies2004.csv")
```

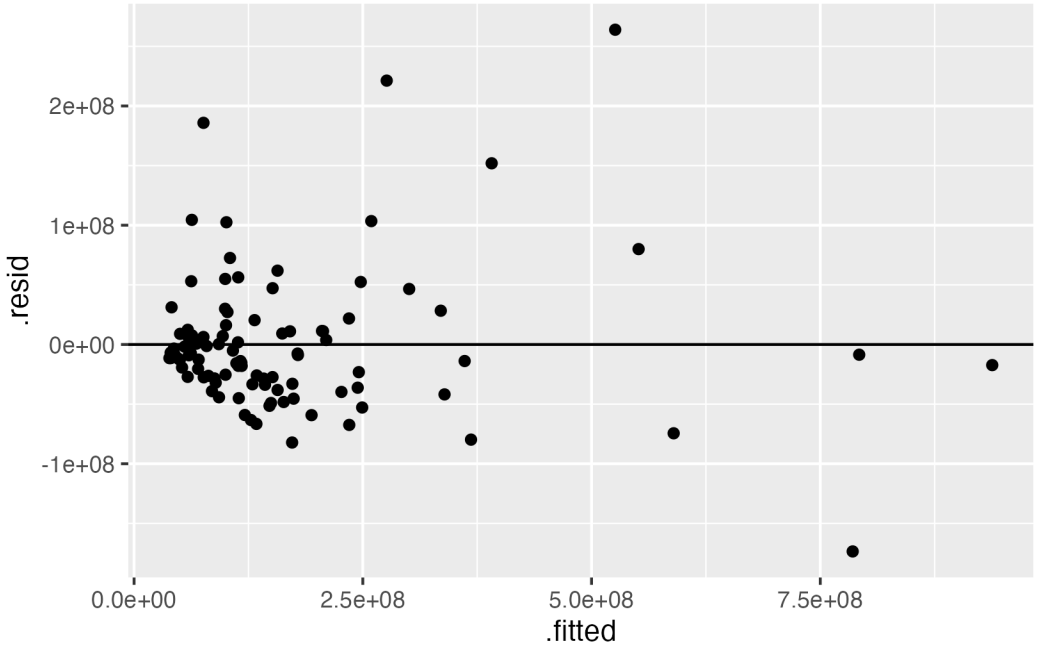
1. Create a scatterplot of `worldwide` against `domestic`. Does the relationship appear to be linear?

2. Run the code below to perform the linear regression of `worldwide` against `domestic`. Interpret the slope and intercept in context.

```
movies_lm1 = lm(worldwide ~ domestic, data = Movies2004)
```

3. Run the code below to create the residual plot with the reference line. Do you think a linear model is appropriate?

```
movies_aug1 = augment(movies_lm1)
ggplot(movies_aug1, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



4. Make a histogram of `worldwide` and `domestic` and describe their distribution.

5. You should have noticed that both variables have a strong right skew. (Your answer should also include roughly where the center of the distribution is). Skew is common in variables involving money, and we often transform them using `log()`. The code below creates those transformed variables in `Movies2004`. Check the distribution of the transformed variables using a histogram. Did the transformation "lessen" the skewness?

```
Movies2004 = Movies2004 %>%
  mutate(
    logworld = log(worldwide),
    logdomestic = log(domestic)
  )
```

6. Create a scatterplot of `logworld` against `logdomestic`. Does the relationship appear to be linear?

7. Run the code below to create the least-squares regression of the log of worldwide gross against the log of domestic gross and call this model `movies_lm2`. Provide interpretations of the slope and intercept.

```
movies_lm2 = lm(logworld ~ logdomestic, data = Movies2004)
```

8. Check the residual plot. Do you have any concerns about the new model?

When you're done, please knit and let Amanda know!

Note: This activity is adapted from the Lab Manual Ch3.5. Some code differs (e.g. using `_` instead of `.` in variable names, using `mutate()`, using `.fitted` in the residual plot) because of my personal code preference, but you can use whichever makes the most sense to you!